Enabling Big Geoscience Data Analytics with a Cloud-based, MapReduce-enabled and Service-oriented Workflow Framework

- An Example with Climate Model Sensitivity Analysis

Zhenlong Li[1], Chaowei Yang[1], Baoxuan Jin[1,2], Manzhu Yu[1], Kai Liu[1], Min Sun[1], Matthew Zhan[1,3]

1. NSF Spatiotemporal Innovation Center, George Mason University, Fairfax, VA, USA
2. Yunnan Provincial Geomatics Center, Yunnan Bureau of Surveying, Mapping, and GeoInformation, Kunming, Yunnan, China
3. Department of Computer Science, University of Texas – Austin, Austin, Texas, USA

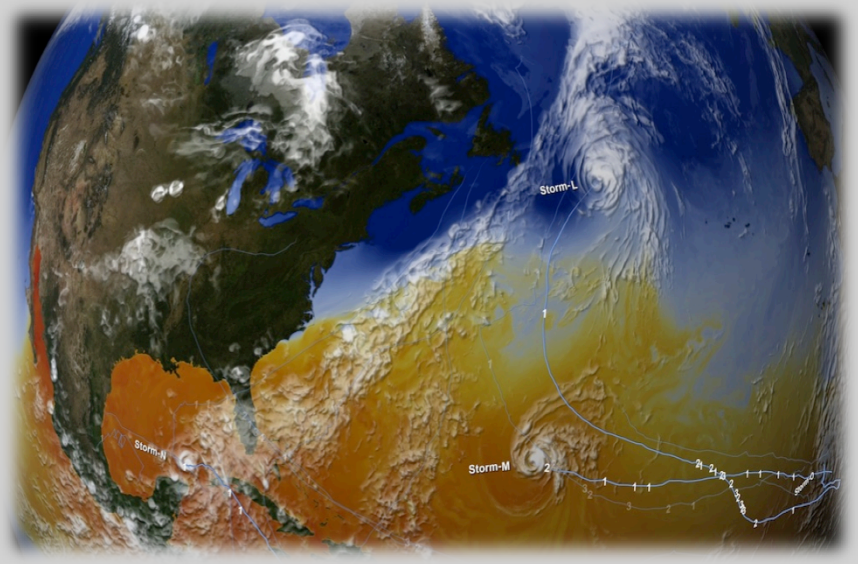
01/07/2015 ESIP winter meeting

Agenda

1. Climate Model Sensitivity Analysis
2. Workflow Framework
3. Result and Demonstration
4. Conclusion

Climate Model Sensitivity Analysis

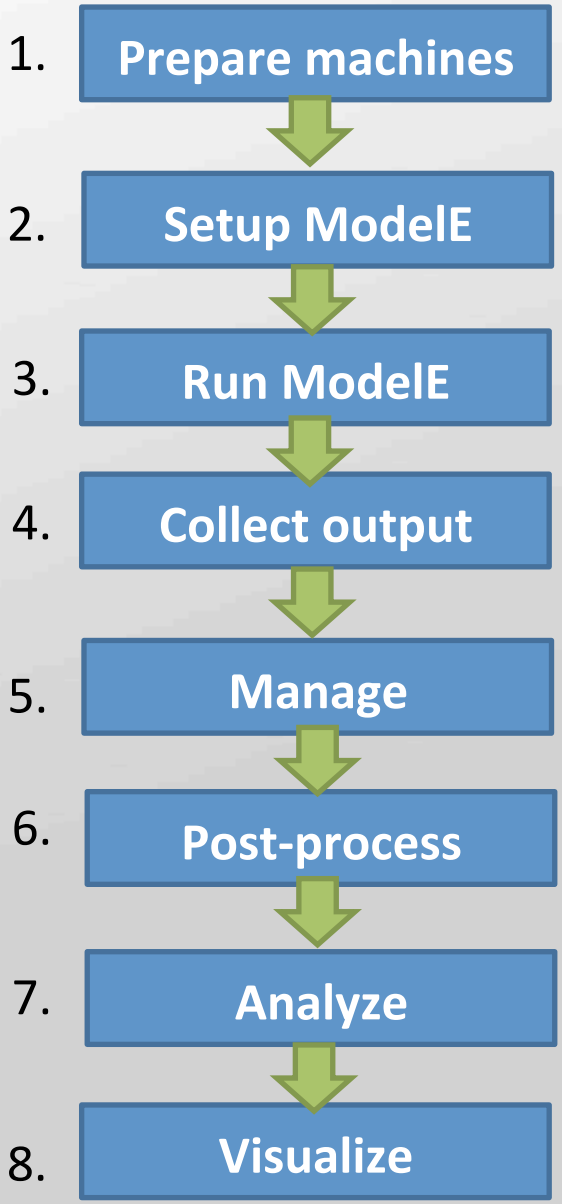
- An essential procedure in climate modeling
- Understand relationships between model input and output
- Run a model many times by sweeping the model input space



An Example

- ModelE: a global climate model developed by NASA Goddard Institute of Space Studies (GISS)
- Test ModelE sensitivity with perturbed physics ensemble(PPE) experiment
- Run ModelE 300 times with different parameter values(PPE-300)
- Conduct PPE-300 many times to investigate how different parameter combinations impact simulated climate

Workflow for the PPE-300



configure 40 machines

install, compile on 40 machines.

300 model-runs

once model-runs are finished

store all data in file system

convert data format, clean data

compute means, analyze correlations of model input and output

Plot, Taylor Diagram etc.

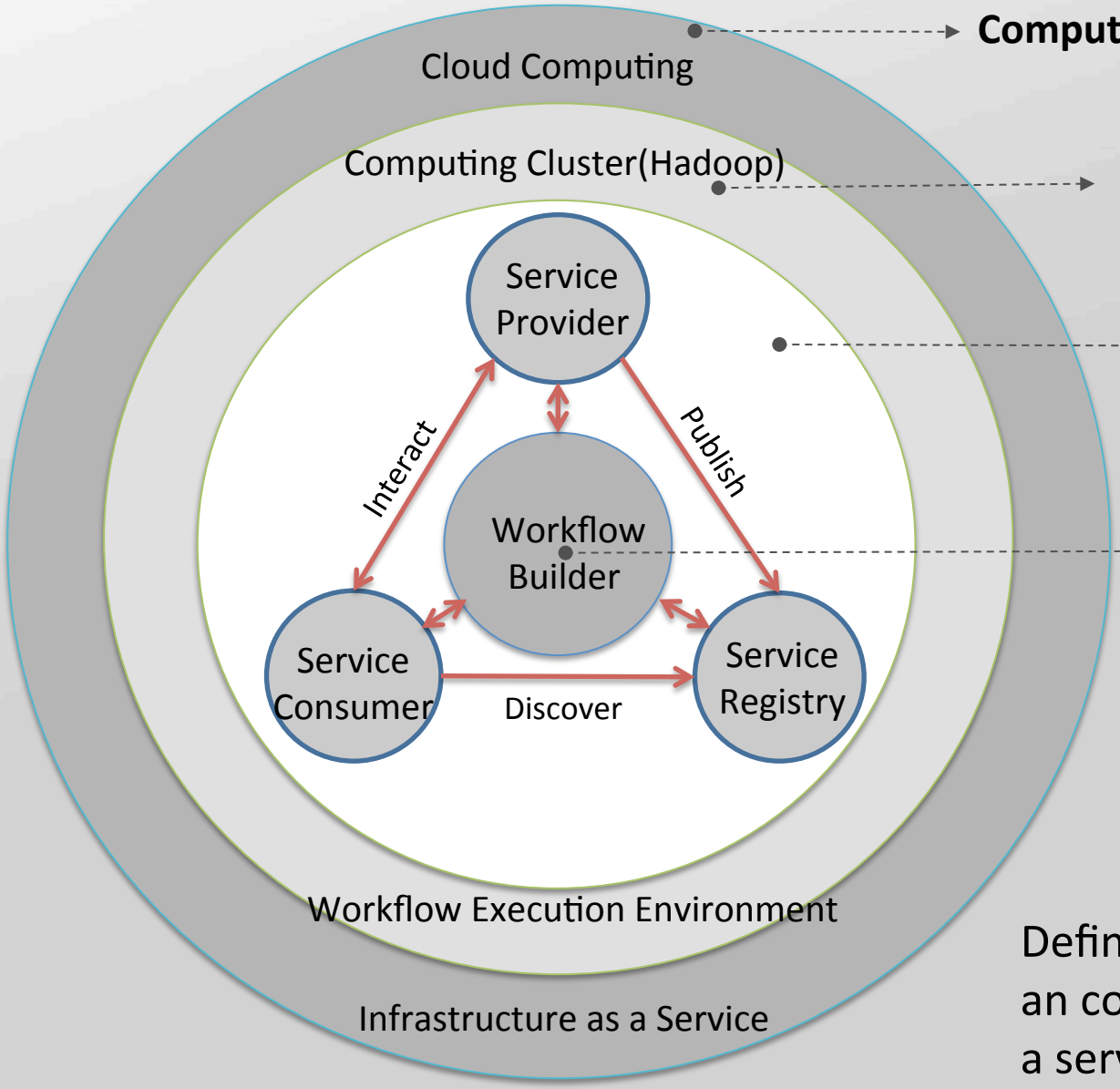
Took two months to finish PPE-300!

A Ph.D. student from GMU collaborated with a NASA scientist

Challenges

- Computing Intensity
  - 300 model-runs →1500 days (4 years)
- Data Intensity
  - One PPE-300 experiment →2.5 terabytes climate data
- Procedure Complexity
  - Multiple steps, different tools, libraries, external services

Cloud-based, MapReduce-enabled and Service-oriented Workflow Framework



Computing resource layer

Processing layer

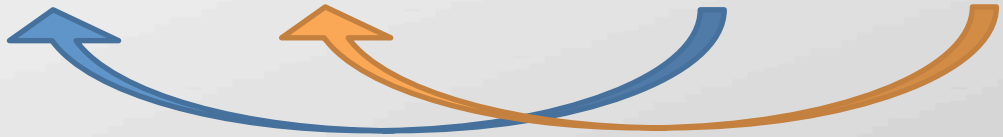
Service layer

User interface

Define each operation in an complex procedure as a service.

Unified Service Model

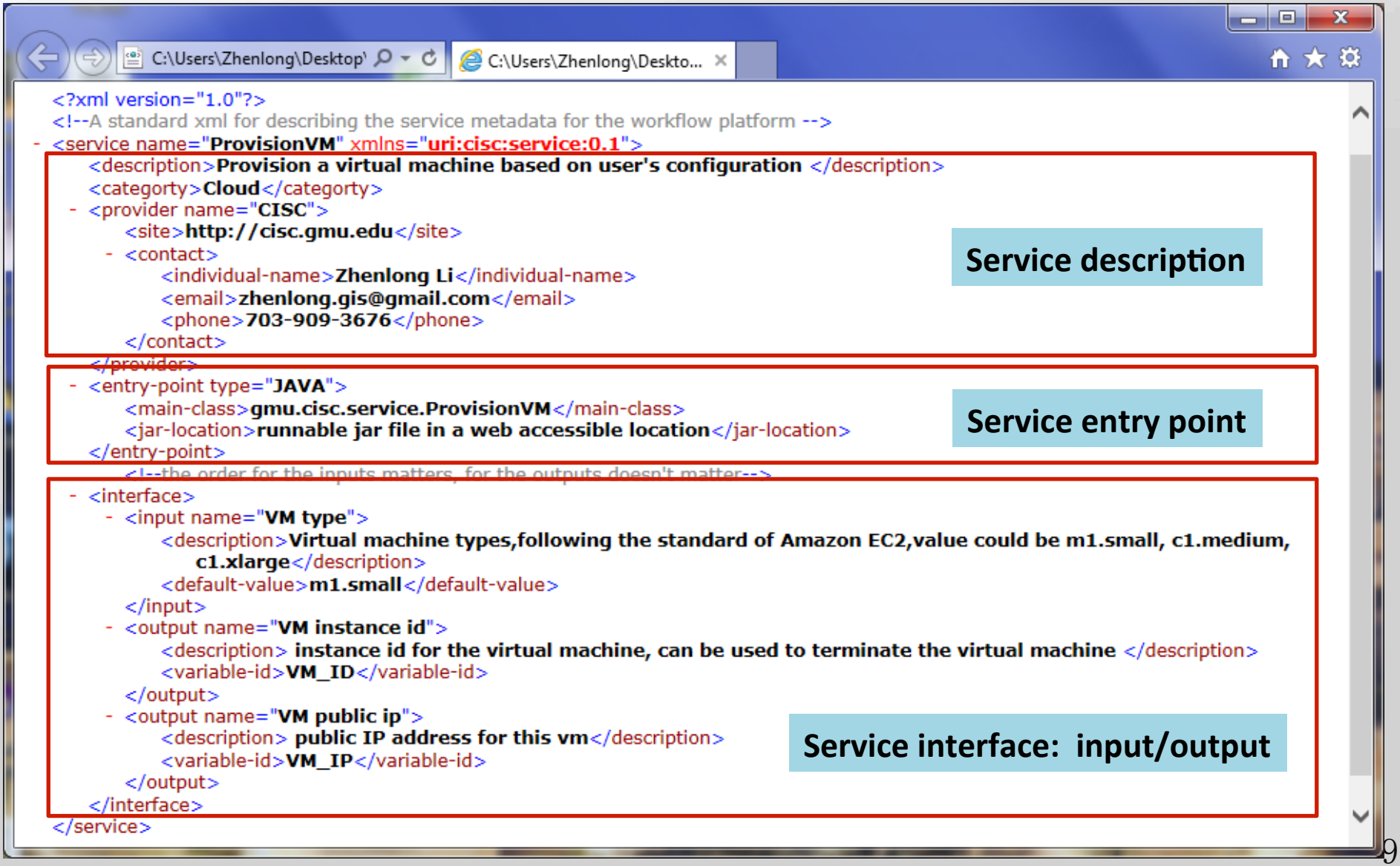
$$\boldsymbol{Service}(DataIn, ParaIn) \rightarrow execute^{\perp}(DataOut, ParaOut)$$



Four service types

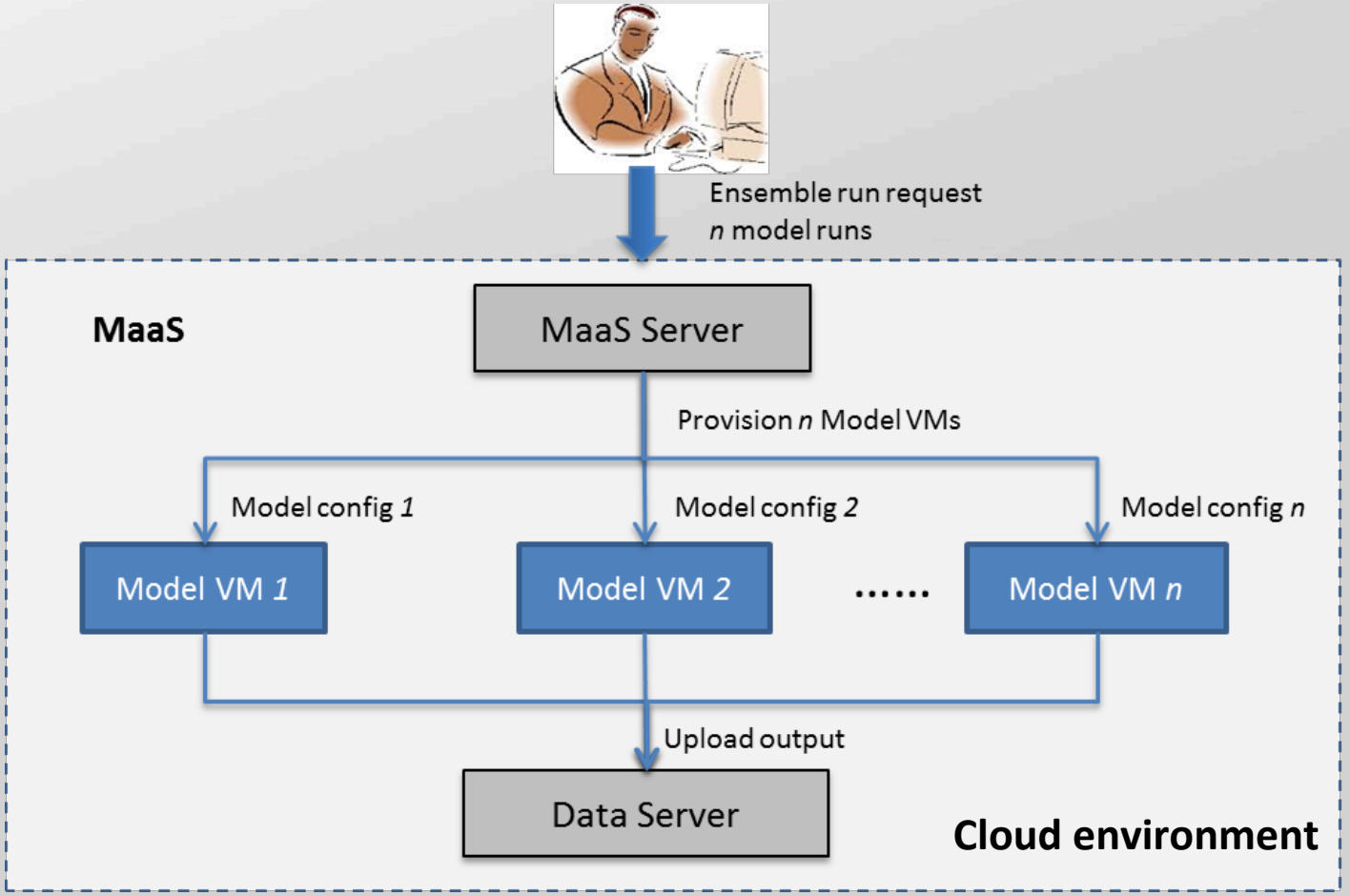
Data Service	• Fetch data from external sources(FTP, OpenDAP) • Publish data
Process/Analytic Service	• Parallel processing(MapReduce) • R script, Python program …
Model Service	• Enabled by MaaS(Model as a Service)
Infrastructure Service	• Start a virtual machine • Start a computing cluster

Standard Service Metadata



Model Service(MaaS)

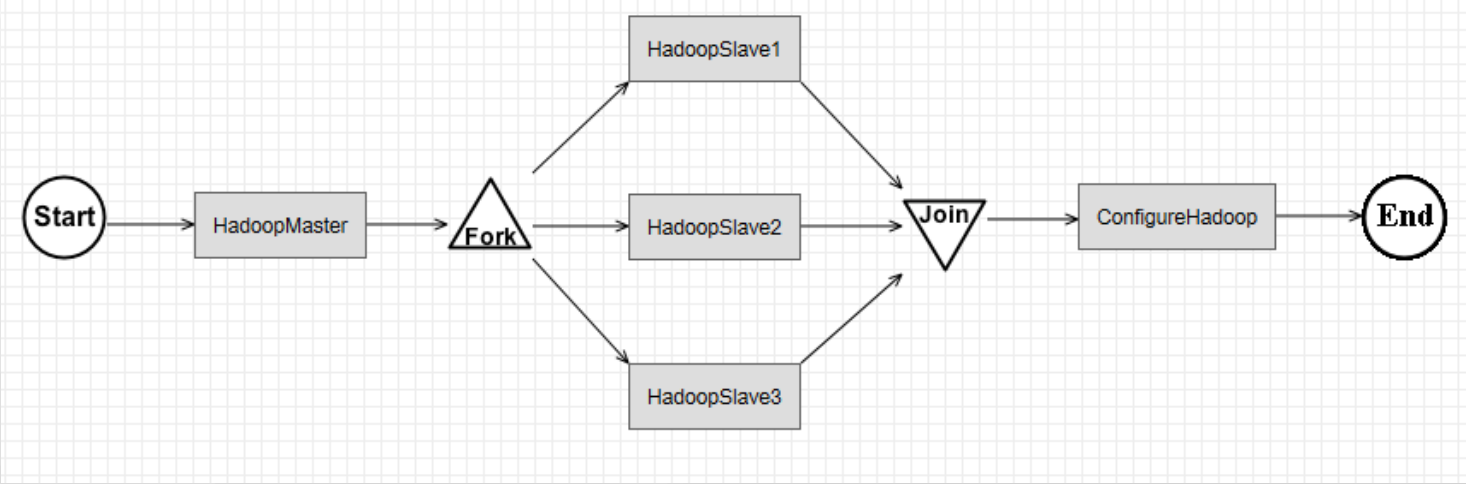
Hundreds of virtual machines with model environment will be started in minutes to run models in parallel



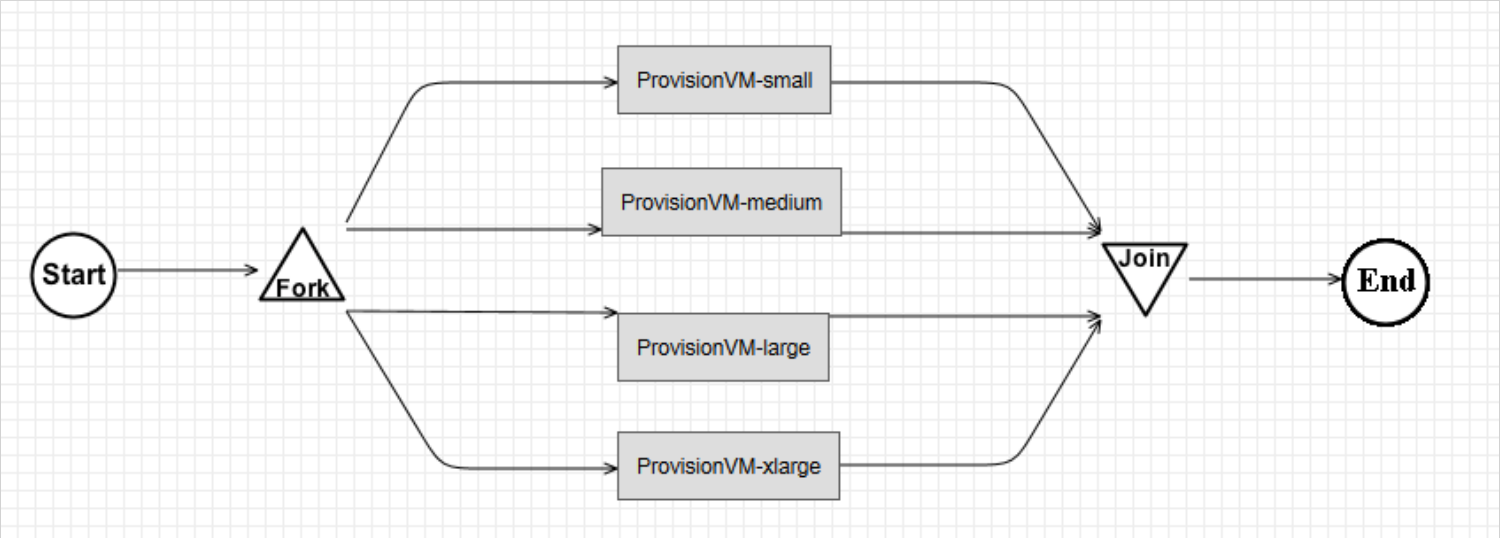
Li, Z., Yang, C., Huang, Q., Liu, K., Sun, M., & Xia, J. (2014). Building Model as a Service to support geosciences. *Computers, Environment and Urban Systems*.

Infrastructure Service

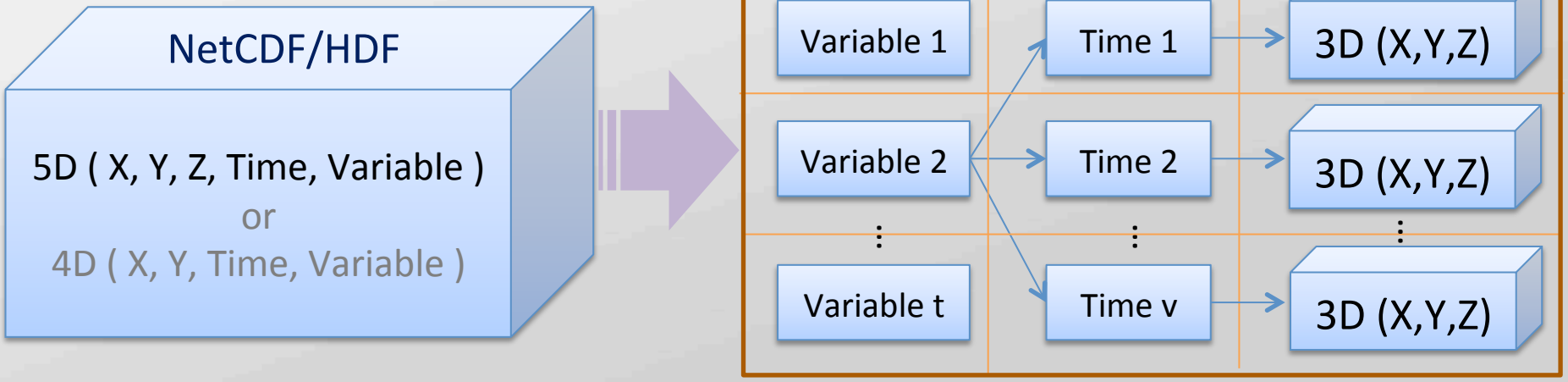
Start a 3-node computing cluster (Hadoop)

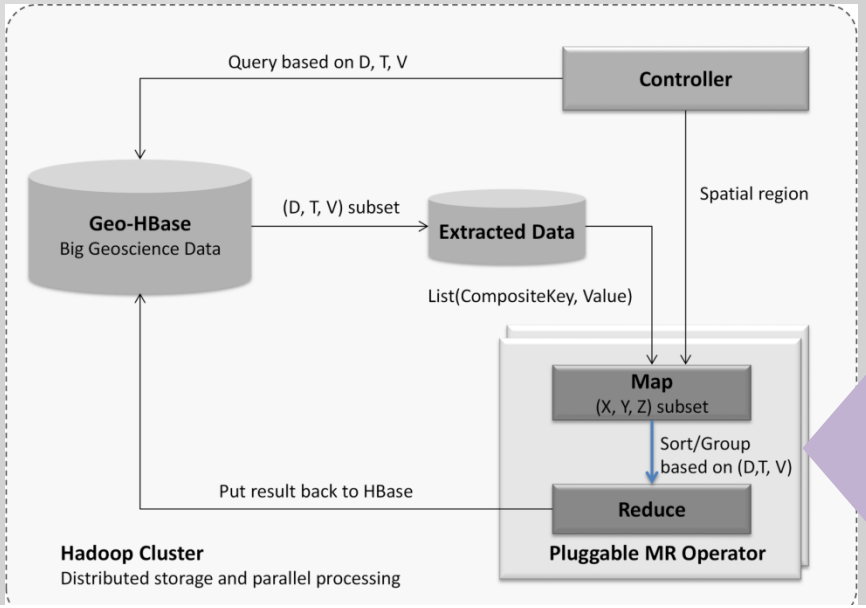


Start four different types of virtual machines in parallel



Big Climate Data Processing with MapReduce





Parallel Analytic Algorithms

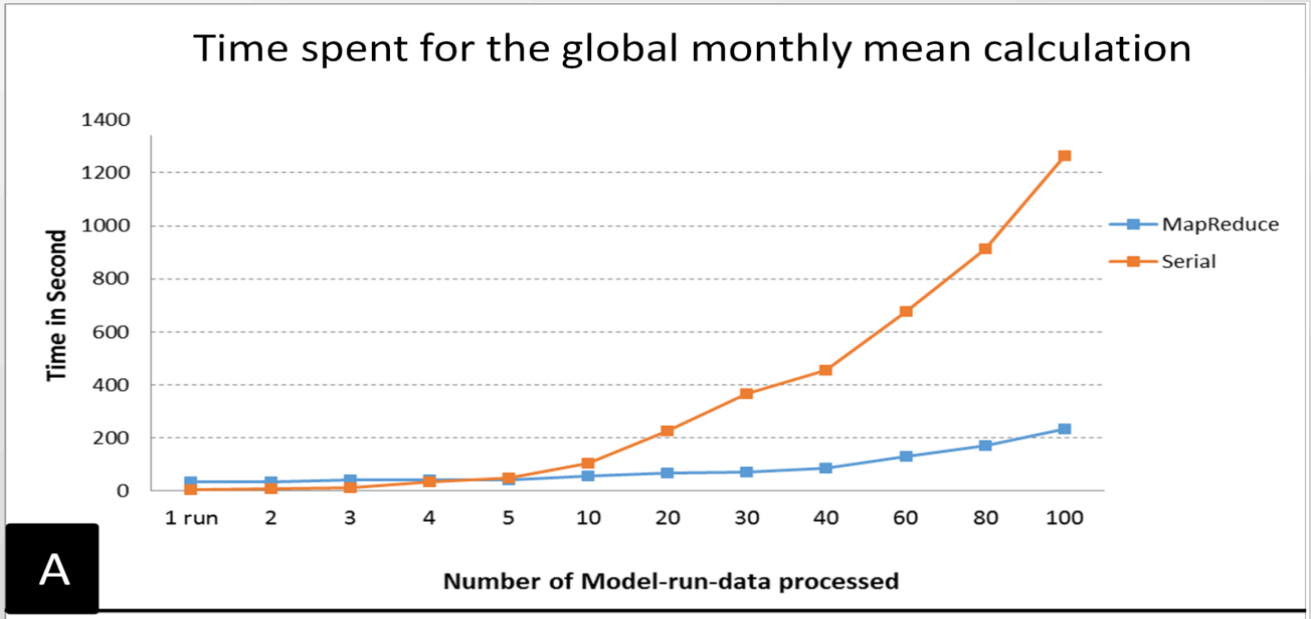
- Max/Min
- Mean calculation
- Anomaly detection
- Interpolation
- ....

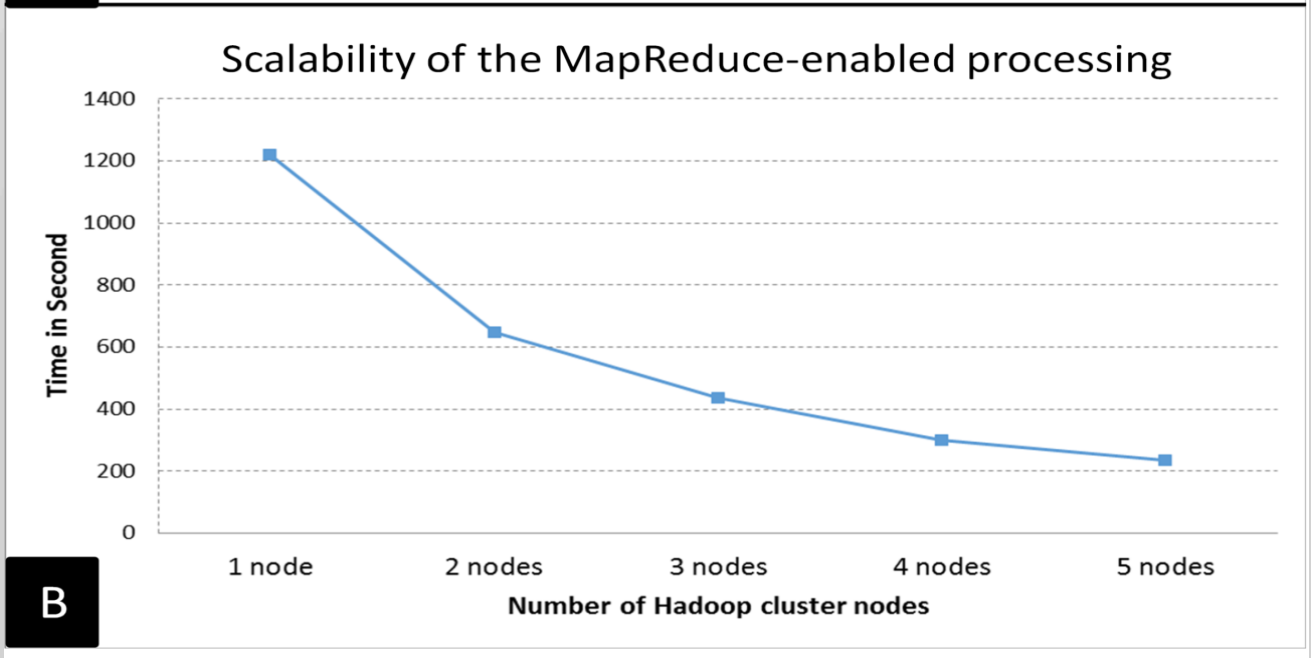
Result: ensemble model run in the cloud

300 model-runs with/without MaaS using the same 40 machines

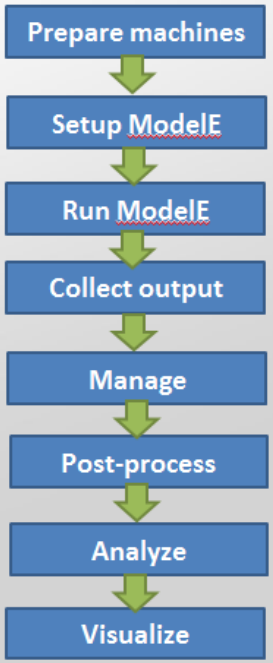
Comparing Item	Without MaaS	With MaaS
Model setup time	~9 days	~10 minutes
300 model-run time	~38 days	~5 days
Resource(CPU) utilization rate	12%	93%

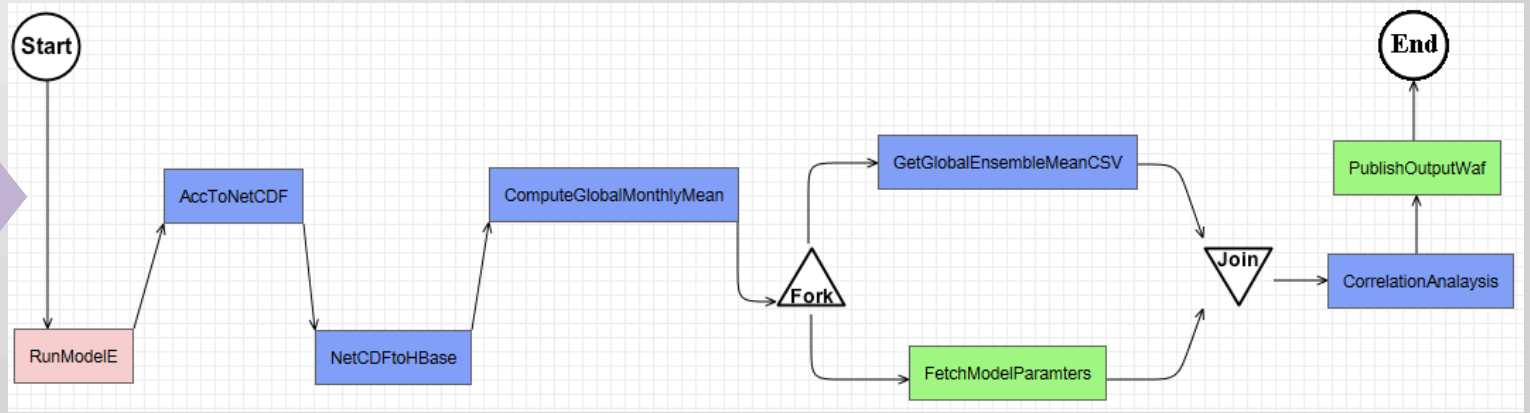
Result: mapreduce-enabled processing





Result: runnable workflow for PPE-300





- PPE-300: finished in 5 days with a few clicks
- Repeat the experiment with minimum effort

Workflow Prototype Demo

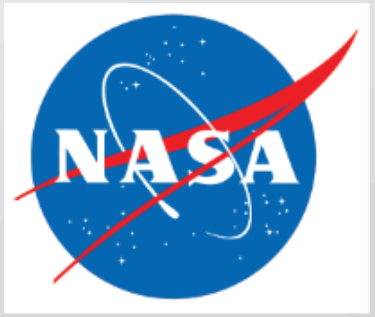
- A video showing the PPE-300 experiment.
- A live demo showing the MERRA data analysis(provision cluster, download data, analyze and visualize).

Conclusion

- **Cloud-based for computing intensity**
  - Dynamically provide computing resources during workflow execution
- **MapReduce-enabled for data intensity**
  - Analyze big climate data in parallel with a highly scalable framework
- **Service-oriented for procedure complexity**
  - Unified Service Model transforms complex experiment into executable workflows

Li Z., Yang C., Yu M., Liu K., Sun M. Enabling Big Geoscience Data Analytics with a Cloud-based, MapReduce-en and Service-oriented Workflow Framework, 2015, *PLOS ONE* (in press)

Acknowledgements



Q&A

Thank you!